Enhancing Clinical Prediction Models through LLM-Agent

Shui Mauser[1], Mohammed Ali Al-Garadi, PhD[1]
[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN

INTRODUCTION

Predicting in-hospital mortality is essential for improving patient outcomes and guiding clinical decision-making.[1] This task is particularly critical in chronic conditions such as heart failure and chronic obstructive pulmonary disease (COPD), where timely interventions can reduce complications and mortality. Electronic health record (EHR) data provides rich data for such predictions, but its complexity due to missing values, heterogeneous data types, and high dimensionality, poses challenges for conventional approaches. Traditional machine learning (ML) models have achieved success using structured data but often require extensive hyperparameter tuning and lack interpretability, limiting their clinical utility.[2,3] Large language models (LLMs) present a means to address these limitations, capable of reasoning over complex inputs and generating human-like explanations. This study integrates an LLM-agent with traditional ML pipelines to improve mortality prediction accuracy and interpretability using structured EHR data from patients with heart failure and COPD.

OBJECTIVES

To predict in-hospital mortality for heart failure and COPD patients, we propose to integrate LLM reasoning with traditional ML methods. We use GPT-4 as an LLM-agent to optimize hyperparameters and perform post-hoc corrections by reviewing patient-level predictions. This hybrid framework combines data-driven ML with contextual LLM reasoning to improve predictive accuracy and explainability.

METHODS

We used data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database (2008–2019),[2,4] including two clinical cohorts: heart failure and COPD. Missing continuous variables were imputed by mean and categorical by mode, with categorical features encoded ordinally. Data were split into training (70%), validation (10%), and test (20%) sets. Four ML models—Random Forest, XGBoost, SVM, and Logistic Regression—were trained via grid search optimization.[5] GPT-4 then refined hyperparameter grids and provided post-hoc prediction corrections with natural language explanations, enhancing both performance and interpretability.

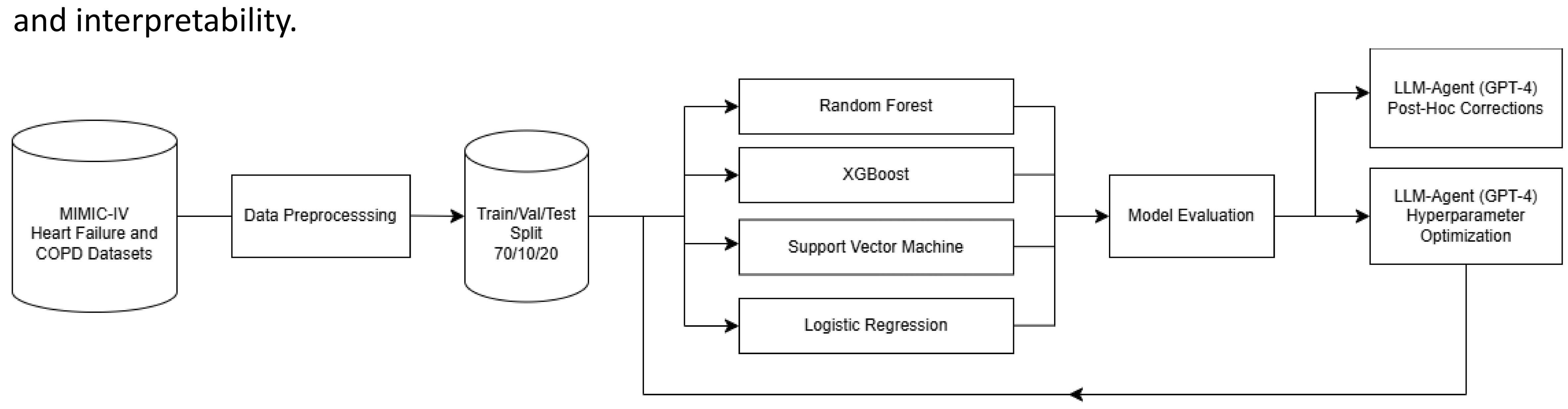

Figure 1 Proposed LLM-agent enhanced modeling framework.

RESULTS

Cohort	Model	Baseline AUC	AUC after LLM-agent Optimization
Heart Failure	RandomForest	.76	.79
	XGBoost	.77	.81
	SVM	.75	.78
	Logistic Regression	.74	.77
COPD	RandomForest	.73	.76
	XGBoost	.75	.80
	SVM	.72	.76
	Logistic Regression	.71	.75

Table 1 AUC performance before and after LLM-agent optimization across clinical cohorts.

DISCUSSION

Integrating an LLM-agent into ML-based EHR modeling improved mortality prediction across all models and cohorts. The LLM's ability to reason over model outputs allowed for more targeted hyperparameter tuning and adaptive post-hoc correction, outperforming standard grid search optimization. Improvements were most pronounced in ensemble methods like XGBoost, suggesting that complex models benefit most from LLM-guided refinement. While the observed AUC increases (0.03–0.05) are moderate, they represent meaningful clinical gains, as small improvements in predictive accuracy can translate to significant differences in patient outcomes. Importantly, these enhancements were achieved without altering the dataset or introducing new features, emphasizing the utility of LLMs as augmentation tools rather than replacements for existing pipelines.

FUTURE DIRECTIONS

The current implementation of the LLM-agent was applied post hoc rather than integrated into live clinical workflows. Future research should pursue prospective validation and evaluate the feasibility of embedding LLM-agent recommendations within real-time clinical decision support systems.

REFERENCES

[1] Soffer S, Klang E, Barash Y, Grossman E, Zimlichman E. Predicting in-hospital mortality at admission to the medical ward: a big-data machine learning model. *Am J Med*. 2021;134(2):227-234.e4.
[2] Chen P, Sun J, Chu Y, Zhao Y. Predicting in-hospital mortality in patients with heart failure combined with atrial fibrillation using stacking ensemble model: an analysis of the medical information mart for intensive care IV (MIMIC-IV). *BMC Med Inform Decis Mak*. 2024;24:402.
[3] O'Reilly D, McGrath J, Martin-Loeches I. Optimizing artificial intelligence in sepsis management: opportunities in the present and looking closely to the future. *J Intensive Med*. 2024;4:34-45.
[4] Johnson AE, Pollard TJ, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10:1.
[5] Brownlee J. How to grid search hyperparameters for deep learning models in Python with Keras. Linea. Published 2016. Accessed October 13, 2025. https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/

VANDERBILT UNIVERSITY
MEDICAL CENTER